Partnership Selection System
Data Mining Proof-of-Concept

By Don McPartland
This paper was presented at the 2002 IRS Research Conference
Internal Revenue Service
LMSB Research

June 12, 2002

Table of Contents

Lists of Tables

Table of Figures

Executive Summary

Introduction

Large and Mid Size Business Division (LMSB) is responsible for administering the tax laws as they affect the country's largest business entities (generally those with assets of at least $10 million). The number of partnerships is growing and they represent an increasing percentage of LMSB workload. Examinations (audits) of returns are a key part of the IRS strategy to ensure tax compliance. It is important that IRS identify returns for examination effectively, efficiently, and fairly. The IRS uses the Discriminant Function (DIF) system to select partnership returns. DIF is developed using results of random examinations to predict audit potential of returns as they are filed. Returns identified by DIF are then manually screened before being sent to Revenue Agent groups for examination. Because partnership DIF is based on data last collected in 1982 and because LMSB returns are a small subset of the total partnership return population, DIF may not be the best way of identifying returns to examine in LMSB. In addition, the manual screening process is costly and somewhat subjective.

LMSB Research identified an opportunity to improve the partnership selection system by applying data mining techniques. The goal was to build a model that would objectively identify returns with significant audit potential without the need for manual screening. The LMSB Research team worked with a representative from Oracle Corporation to build the model (referred to as Model 22). It successfully automates the process by which returns are selected for examination or accepted as filed. It was built only for the LMSB portion of the partnership workload; however, the methodology used could be applied to the whole partnership population and to other types of tax returns as well.

This report details the development of Model 22, including sample selection, data collection and preparation, data analysis, findings and recommendations.

Business Objective

In this project, the business objective of Large and Mid Sized Business Division (LMSB) is to develop and test an automated return selection system that identifies potentially non-compliant partnership returns by replicating judgments made by experienced revenue agent classifiers.

Data Gathering and Data Preparation

During Processing Year (PY) 2000, there were about 2.0 million partnership returns filed. Of these, about 50,000 were LMSB returns (had assets of at least $10 million). A stratified random sample of 1,790 returns were selected from the LMSB population and classified by experienced revenue agents. Returns were categorized as "selects" if they appeared to have significant audit potential (exact criteria is left unspecified here) or "accepts" if they did not. Of the 1,790 returns classified, 682 returns were identified as "selects." Partnership Technical Advisors, employees most knowledgeable about partnership issues, reviewed the determinations made by the revenue agents. The Technical Advisors changed 177 of the "selects" to "accepts." These returns were not used to build the model.

A limited number of data fields are transcribed as returns are processed. This data is contained in the Business Returns Transaction File (BRTF). This electronic data was obtained for all LMSB partnership returns. For the sample, the "select/accept" designation was appended to the BRTF data for the sample returns. Data was reviewed for accuracy and completeness and to identify unnecessary or redundant information.

Model 22 Development

The next step was to determine if it was possible to accurately categorize returns as select/accept using only the transcribed data from the BRTF. Darwin software, from the Oracle Data Mining suite, was used in

the development of the partnership model. Darwin is a data mining tool that builds predictive models by finding meaningful patterns and correlations in the data. It contains a model seeker wizard that automatically runs multiple models using different algorithmic data mining approaches—Neural Networks (Nets), Classification and Regression Trees (C&RT), Memory-Based Reasoning (Match), and Clustering (Cluster). This comprehensive array of techniques increases modeling capabilities and accuracy. After comparing the results from each technique, Oracle determined Darwin tree (C&RT) produced the best fit for this project's data and business objective.

Data was divided into three categories (Train (70%), Test (15%), Predict (15%)). The largest data group was used to develop fifty different C&RT models that were tested using the second data set to improve accuracy. The data set was then used to evaluate each model. In the end, Model 22 was found to be the best predictor of the select/accept categorization. Model 22 contains nineteen nodes or decision points.

Model Application

The completed model was applied to all LMSB partnership returns for processing years 2000 and 2001. Each record of the BRTF was scored (except for the sample PY 2000 returns used in model development) and identified as a "select" (should be examined) or "accepted" (should not be examined). A corresponding confidence level was also computed. The following table reflects the results:

PY	Number of Returns	Number of selects by Confidence Levels				
		70-79%	80-89%	90%+	Total	Percentage of Selects
2000	53,528	9,330	817	965	11,112	20.8%
2001	60,142	10,232	2,296	1,187	13,715	22.8%

Model 22 was compared with DIF. At the 70% confidence level, DIF and Model 22 agreed 54% of the time. Model 22 identified 8% more returns as "selects" than DIF.

Model Tests

Several tests were conducted to determine how well the model works; this process is on going. First, the Partnership Technical Advisors reviewed the decision tree. They found that the rules used in Model 22 were intuitively reasonable. Second, the model was run on the 177 returns reclassified as "accepts" after being originally classified as "selects" by the less experienced classifiers. Model 22 agreed with the Technical Advisors 94% of the time. Finally, revenue agent classifiers found such a high percentage of agreement with Model 22 "selects" from PY 2001 returns that manual classification was discontinued. While these tests may not validate the model, they serve as an indication of how well the model predicts productive returns. The model is now being run on prior year returns and will be compared with actual examination results. Returns selected using Model 22 are being tracked and audit results will be available for these returns in 12-18 months.

Conclusion

Research West in concert with Oracle Corporation successfully developed and delivered an automated scoring model that selects potentially non-compliant partnership returns by replicating judgments made by experienced classifiers. Data mining is a useful approach in developing risk assessment and return selection models.

Recommendations

Oracle and the LMSB Research team made the following recommendations:

1. Begin using Model 22 as an adjunct to current return selection processes. Select returns first when DIF and the Model agree. If more returns are needed, use Model 22 selects down to the 70% confidence level.
2. Continue to validate the accuracy of Model 22 using historical audit results and the results of returns selected using the model.
3. Use additional data (from prior examinations) to build a model that ranks returns by audit potential and can be used to estimate levels of voluntary compliance.
4. Conduct a cluster analysis of the entire partnership population to determine homogeneous populations that can be used to build models for all partnership returns.

I. Business Objective

In this project, the business objective of Large and Mid Sized Business Division (LMSB) is to develop and test an automated return selection system that identifies potentially non-compliant partnership returns by replicating judgments made by experienced revenue agent classifiers.

II. Research Objective

Determine if data mining methods can be used with information from the Business Returns Transaction File to develop a model that meets the business objective.

III. Introduction

LMSB Research, in concert with Oracle Corporation, completed the development of a partnership return-scoring model to select partnership returns with a high potential for examination. This was an opportunity for the Oracle Corporation to demonstrate its ability to solve an LMSB business problem through the use of a statistical process known as data mining.

The model was developed with the intent of reducing the amount of resources expended on classification by materially increasing the select rate of partnership returns pulled by the service center for classification. The model automates the classification process by replicating the work of expert classifiers and complements the current DIF selection process.

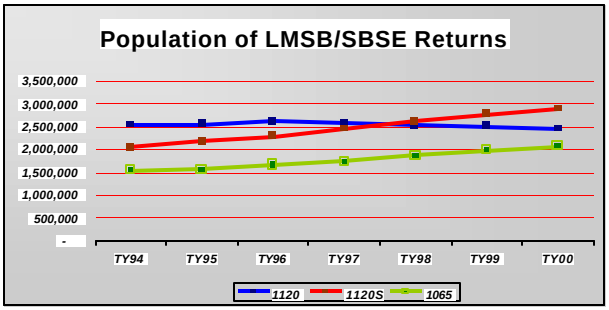
The Partnership Project was completed in two phases. Phase I involved data gathering, data preparation and model development by the staff of LMSB Research and Oracle Corporation (April 2001 through January 2002). Phase II entails ongoing model testing (February 2002 to present).

The completed model was applied to the LMSB Business Return Transaction File (BRTF) for Processing Years (PY) 2000 and 2001. Each BRTF record was categorized as a "select" (should be examined) or "non-select" (accepted as filed) and assigned a confidence level indicating the likelihood that the categorization was correct.

Partnerships are required to file Forms 1065 to report income and loss. Income is not taxable to the partnership but income (and losses) flow through to the returns of the partners. Partners may be individuals, trusts, corporations, or partnerships (tiers). Partnerships may be engaged in a trade or business or simply pass through income from investments.

LMSB Division partnership returns have reported assets of at least $10 million. The Small Business/Self Employed Division (SB/SE) works returns with less than $10 million in assets. Partnership return filings for both LMSB and SBSE have steadily grown since 1994. Partnership population increased from 1.5 million in PY 1995 to 2.1 million in PY 2001-up by 33.8% (refer to Figure 1).

Figure 1 Growth Comparison of the Business Entities' Filings



While the partnership population has grown steadily, the partnership audit coverage level had steadily declined. In FY 2001, the coverage level was .25%, down from .5% in FY 1994. In FY 2001, the coverage level for partnerships compared to the other entities is shown in Table 1 below.

Table 1 FY 2001 Total Returns Filed and Coverage Levels of Selected Business Entities

Type of Business Entity	Number of Returns Filed	Coverage Level
Corporations	2,453,000	.95%
Subchapter S Corporations	2,887,100	.43%
Partnerships	2,066,800	.25%
Fiduciary	3,528,900	.20%

For more than 30 years, the Internal Revenue Service has used the DIF system to score tax returns for audit potential based on the relative probability that a given return contains a material error affecting the amount of tax due. DIF formulas are developed using data from Taxpayer Compliance Measurement Program (TCMP) examinations of a random sample of returns. The DIF formulas for partnership returns were developed using the 1982 TCMP data. Tax laws and taxpayer behaviors have changed since these examinations. In addition, DIF is used to score the entire partnership population and may not be valid for the smaller LMSB sub-set of partnership returns.

Revenue agents manually screen LMSB returns identified by DIF as having high audit potential along with returns meeting certain other criteria before being sent to the field for examination. This is a resource intensive process and is somewhat subjective.

A significant increase in the partnership filing population, a growing compliance risk, and a desire to increase examination coverage of partnership returns created a need to improve the effectiveness and efficiency of the partnership selection system. At the same time, reducing subjectivity and improving the accuracy of classification means increased fairness and reduced burden to compliant taxpayers.

One method to develop a scoring model would be to use examination results from closed cases. This is the method that was used to develop a scoring system for LMSB corporations. However, the percentage of partnership cases examined is very low and cannot be used to approximate a random sample. Additionally, it is thought that the method by which returns were selected for examination contain specific biases that cannot be overcome. Therefore, the results of previous examinations were not used for the partnership model. Instead, the model is intended to reproduce human judgments made by classifiers.

To develop the model, a random sample of LMSB returns was selected, classified, and reviewed. Returns were categorized as selects or non-selects. This data was combined with fields from the return that are transcribed and included on the BRTF. A data mining approach was then used to develop the predictive model. This work was done in conjunction with the Oracle Corporation. The Oracle data mining suite uses Classification and Regression Tree (C&RT) algorithms to extract multiple cause-and-effect relationships from a data set on a specific dependent variable. It extracts significant *attributes* (features) and *intervals* (ranges). Data mining is a methodology used to autonomously interrogate a database for patterns and clusters. Data mining is based in part on statistics and a field of artificial intelligence designed to emulate human perception known as *machine-learning*. Unlike traditional data analysis programs, data mining tools perform the analysis automatically and formulate their solutions in graphical decision trees or set of rules.

The data mining process delivers returns to the classifier with a greater potential for examination. It is anticipated that the model will minimize the number of non-productive returns to be classified while reducing the amount of time and resources needed to classify and select the required workload for examination. (Fewer returns will need to be classified to arrive at the desired number of returns to be examined.)

By automating the classification process the Service will be able to classify the entire partnership return population and predict returns for examination based upon significant relationship of key line items on Form 1065. This will allow the Service to focus resources on the more productive returns and reduce taxpayer burden.

IV. Data Preparation

A. Sampling Design

Sample Frame

The sample was extracted from PY 2000 partnership returns with a TC 150 posting (original return), and total assets greater than or equal to $10 million (55,288 returns). The sample was drawn using a 95% confidence level, with an acceptable error level of plus or minus 5%.

Sample Size

A stratified random sample method was used to obtain an equal representation of partnership returns in each of the five LMSB industry designations and each asset class (see Table 2 on the next page). The sample size is 1,790 returns. Three hundred returns were drawn from each of the five LMSB industries and two hundred ninety returns from the invalid or unclassified industry[1] groups. The sample size should be eighteen hundred except for the fact that the population in the asset class of greater than $250 million in the unclassified industry category only has a total of 50 returns, 10 short of the desired 60.

Table 2 Number Partnership Returns by LMSB Industry (PY 2000)

LMSB Industry	Returns	Percent of Total	Sample
Financial Services	17277	31.24%	300
Natural Resources	4245	7.67%	300
Communication, Technology & Media	3080	5.57%	300
Retail, Food & Healthcare	4520	8.17%	300
Heavy Manufacturing & Transportation	25085	45.37%	300
Invalid & unclassified	1081	1.95%	290
Total	55288	100%	1790

Source: PY 2000 Business Return Transaction File (BRTF)

Activity Codes were not considered in the sample design because the codes are based upon gross receipts and number of partners and are not considered useful in categorizing LMSB partnerships.

Partnership Activity Codes

481 with 10 or less partners and gross receipts under $100K
482 with 10 or less partners and gross receipts $100K and over
483 with 11 or more partners irrespective of gross receipts

[1] An Invalid return has an IRS industry code that does not convert to a valid North America Industry Classification System (NAICS) code. An unclassified return is a return with an IRS Industry code value of either 999000 or 999999.

Instead, the sample was stratified by industry using the following asset categories to make the distinction between the operating divisions and to align the data with established corporate asset classes.

$10 to $14.9 Million

$15 to $49.9 Million

$50 to $99.9 Million

$100 to $249 Million

>=$250 Million

B. Data Gathering and Development of Check Sheet

A data-gathering instrument was developed specifically to capture information during the classification process. LMSB Research, with the assistance from revenue agents and the Partnership Technical Advisers, prepared a classification check sheet that captured line item issue(s) and the estimated percentage or dollar amounts corresponding to those issue(s). The check sheet replicates the line item information on Form 1065 with the addition of an estimated dollar amount of adjustment or percent of item adjustment and the reason for selection columns.

The check sheet used in the classification has three distinct parts:

- Part I contains line items from the four pages of the Form 1065
- Part II contains the instrument that identifies specific issues developed by Technical Advisors and Revenue Agents
- Part III of the instrument includes the remarks/ comments section and the classifier select/non-select determination (ranking).

After the sample of returns was selected, labels with name, address, and TIN were generated, attached to the check sheet and associated with the returns. Additionally, a BRTF transcript for the two immediate preceding years was generated and associated with each return. It was anticipated that this additional information would assist the classifiers in identifying issues and estimating adjustment amounts, however, feedback from the classifiers indicated this was not as beneficial as originally thought.

C. Classifier Selection and Training

Industry Directors were asked to provide the best available revenue agents to classify the sample returns. An analysis of sample size and estimated time to classify returns was made to determine the number of classifiers needed and the duration of each classification session. Operations Support coordinated the requisitioning of sample returns, space and support at the Ogden campus and the selection of revenue agents for the classification details. Prior to each of the three classification sessions, the Partnership Technical Advisers from LMSB, SB/SE and the Research staff conducted a one-day orientation and training workshop for classifiers. The Technical Advisors remained available for assistance for the duration of the classification sessions. This training provided for a better understanding of the project objective, a discussion of significant issues such as disguised sales, “mixing bowl” transactions, book and capital accounts, partnership allocations, and real estate partnerships and general partnership classification procedures. Instructions were provided on “ranking” classified returns as Category A, B, or C. Original ranking definitions were:

- Category A (selects)- Highest examination potential (specifics omitted here);
- Category B (eliminated) - returns with a significant audit item but do not necessarily have an adjustment;
- Category C (non-selects)- returns that appears to be accurate and complete with no potential adjustment.

Category B was eliminated after the first classification session because, given workload and resource considerations, returns in this category would not be selected for examination. This made the ranking determination more straightforward for the classifiers.

There were 682 returns classified during the first session. Of these 682 returns there were 175 “A’s”, 108 “B’s”, 399 “C’s”. After a quality review by the Technical Advisors, 28 of the ”B’s” were reclassified to “A’s” and the remaining returns were changed to “C’s”.

D. Classification Efforts

Not all returns were available for classification; some were in Statistics of Income (SOI), awaiting transcription of K-1 data, or were charged out to the field. A total of 3,636 returns were ordered to meet the desired sample of 1,790. Of these 2,086 returns were received. However, only 1,760 met the correct stratification criteria and were used in model development.

Classification was done in three sessions. The original plan was to complete classification in two sessions. However, due to the limited number of experienced agents available and a delay in receiving the requested returns an additional classification session was needed. The first classification session was held July 10-12, 2001 in Oakland because of space limitations in Ogden. Ten revenue agents classified 694 returns. The second classification session was held August 6-9, 2001 in Ogden. Thirteen revenue agents classified 918 returns. The last classification session was held on September 18-20, 2001 in Ogden. Seven revenue agents classified 474 returns.

A photocopy of the check sheet and the first four pages of the return were made for each classified return. This was done for potential future analysis since selected returns were forwarded to Operations Support for field distribution and non-selected returns were returned to files.

Upon completion of the classification sessions, a database of classified returns and their rank was created and used for model development.

E. Quality Review

The Partnership Technical Advisers conducted a one hundred percent review of returns ranked Category “A” and a random sample of returns ranked Category “C”. The goal was to conduct a one hundred percent review of the Category “C” returns; however, this was not feasible due to time constraints and multiple priorities placed on the technical advisers.

Of the 1,760 classified returns, there were 635 (36%) selects and 1,125 (64%) non-selects. After review, 177 of the original 635 selects were re-classified into non-selects. These reclassified returns were not used for model development. The final development file used contained 1,583 returns, 458 selects (29%) and 1,125 non-selects (71%).

F. Data Needs

Primary Data

1. After initial classification of the stratified sample, a classification results database was created containing data from the Research partnership classification check sheets.
2. A database was created containing all the information from the check sheet including line item issues estimated dollar adjustment or estimated percentage of adjustment, and ranking. The above information was extracted and compiled into an access database, and
3. A database was created containing the PY 2000 Business Return Transaction File (BRTF) and the appended ranking of each entity. This third database was used in model development.

Secondary Data

1. PY 2000 – 2001 Business Master File/Business Returns Transaction File (BMF/BRTF) – This database file is a special extract, which contains selected account-related items from the Business Master file and the complete Business Returns Transaction File.
2. Audit Information Management System (AIMS) Closed Case Database File (CCDF) – AIMS CCDF databases contain information about Examination results for Fiscal Years 1992-2001.

V. Model 22 Development

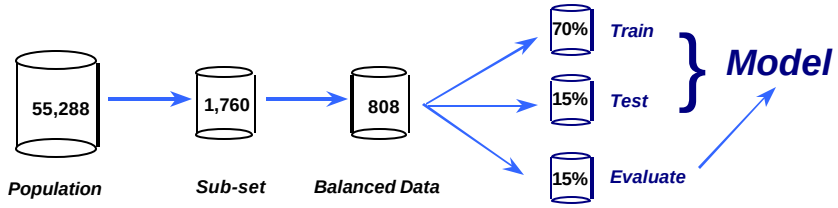
A. Data Validation and Transformation

Each record contains 127 variables. During the data validation process, all data fields were analyzed and all non-relevant fields were removed. There were also additional fields created for asset class, rank, rank2, and rank3. The asset class field was derived from the categories of total asset field. Rank field contains the original ranking (with Category B). Rank2 field contains the audit trail of the ranked B returns that were changed to either A or C. Rank3 contains the final ranking given to each of the returns. The fields were reduced from 107 to 70 based on some of the fields not being relevant for the modeling process.

Blanks and missing values were also analyzed. Oracle recoded real blanks as null and user blanks as zero.

Because of a high percentage of non-selects (71%) compared to the selects (29%), a balanced data set was created. The purpose for balancing the data was for the model not to over fit the pattern recognition on the non-selects. After balancing the sample data, a mutually exclusive set of randomized records from the original set of all classified returns was created for modeling purposes. The data was divided into Train (70%), Test (15%), and Predict (15%). The train and the test data sets were used in building the model and the 15% predict file was used in evaluating the model performance.

Figure 2 Schematic Diagram of the Data Transformation for Modeling



B. Model Building & Analysis

Darwin data mining software was used in the development of the partnership model. Darwin is a tool that builds classification and predictive models by finding meaningful patterns and correlation in the data. Darwin contains a model seeker feature that automatically builds multiple models using different data mining techniques—Neural Networks (NN), Classification and Regression Trees (C&RT), Memory-Based Reasoning (K-Nearest Neighbor), Clustering (K-Means) and Logistic Regression. This comprehensive array of techniques increases the ability to create accurate models based on the problem that is being solved. For partnership model development, Oracle used these various techniques. After comparing the results from each technique, Oracle determined C&RT produced the best fit for this project's data and business objectives.

Darwin's modeling techniques are based on two different types of learning, Supervised Learning and Unsupervised Learning. Supervised Learning is a set of modeling techniques used when there are sufficient historical records called targets for the data mining algorithms to learn from, and find and detect a pattern. Unsupervised Learning is another set of data mining techniques used when there are insufficient historical records for the algorithms to find and detect any pattern. This technique has the capability to take data and create groups based on similarities without the use of any historical target field from which to find a pattern.

In the development of the Partnership Model (Model 22), a supervised learning method was used and a target field was identified. The target field used for this modeling effort was called "rank3" which was used to define whether the tax return should be selected for an audit or not selected. "Rank3" is a binary field with values A for select and C for non-select and was the classifier's best judgment of the return's rank. Therefore, the models developed were based on group patterns detected during classification of returns. The predictor fields used to identify patterns based on the target field were the 70 line items of the Form 1065. Using Supervised Learning Techniques with "rank3" as the target field and the 70 line items of Form 1065 as the predictor fields, Oracle was able to build about 50 development models using C&RT and 2 models using Neural Networks. After testing all of the models, it was decided that Model 22 provided the best overall performance and it was chosen for the production model to be used as part of the classification process of the PY 2001 partnership returns.

The technique within Supervised Learning that was used to develop Model 22 is called Classification and Regression Trees (C&RT). This technique is used to generate decision rules for making binary (0,1) multi-class (1,2,3,4) and regression (16, 45, 67, 89) type predictions and classifications. The data starts as one heterogeneous mass and through continuous splits of the data, refines the records into smaller and smaller segments or mutually exclusive groups. Each split creates branches or paths that are used to create groups. The splitting occurs until no more splits can be created giving you segments or groups, which are as homogeneous as possible. At the end of each branch is a set of leaves or nodes that belong to a specific value of the target field. Once this tree like shape has been produced, it is important to prune the tree to account for over fitting and ensure the lowest error rate. The end result is human readable rules and tree-like diagrams based on the splits. Using Darwin's C&RT technique, the data splits can be viewed in a tree-like diagram or rules format.

After the data is properly prepared for modeling by taking the source data set and splitting it into equal thirds or mutually exclusive subsets of the original source dataset as described in the section "Data Transformation" above, there are three main steps that are necessary in developing a model. The first step is to train the model. Training a model means to create the mathematical representation of the data called the model which will be able to examine the historical records that are associated with each of the target values for that given record. This will produce a model that will find patterns and relationships that will be able to predict the values of the target field for unclassified data with some degree of probability. The next step is to test the model for under fitting (to general), over fitting (to specific) or any idiosyncrasies that may have occurred during the training phase. The final step is to test the models overall performance to see how well it will work in making predictions when applied in a production environment to unclassified data. During the testing phase the development model created during the training phase is used to make predictions against the historical data while ignoring the actual value of the target field for each of the records. Once this has been done, the actual values and the predicted values are compared to each other for overall model accuracy and performance.

Darwin's C&RT uses two diversification algorithms to split the data into heterogeneous groups. The algorithms are called Gini and Entropy. These algorithms use orthogonal lines to decide how the homogenous populations are split into heterogeneous groups. In the building of Model 22, the Gini algorithm was used for splitting. The cost algorithm was used to control model over fitting by pruning the branches of the tree back and creating different pruned versions of the tree called sub-trees along with their associated error rate. The error rate or misclassification cost is the overall error of classifying a record to one group when it really belongs to another. Once each sub-tree was tested for its individual error rate, the sub-tree with the lowest error rate was chosen to test the overall performance of the models accuracy in classifying returns as either select or non select and creating type I and type II errors. An operational decision was made to minimize type II error and maximize type I error for all returns that were not classified by the model correctly. Type I error is a type of misclassification where a return is

identified as a select but could potentially be a non-select. Type II error is a type of misclassification where a return is identified, as non-selects but could potentially be a select. This decision was made to err on the side of Type II in order to reduce the opportunity cost (revenue lost) involved for not pulling and reviewing the cases that were predicted to be non-selects cases but have some probability of being select cases.

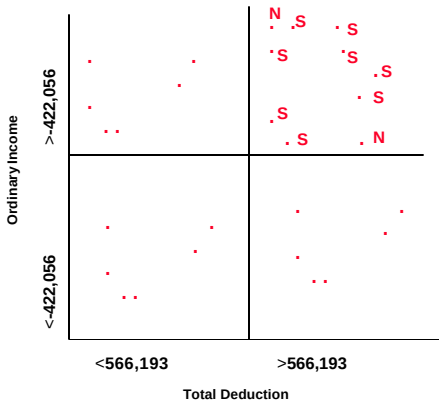
Model 22 contains nineteen-tree nodes, nine of the nodes had decision rules that produced select classifications and ten of the nodes had decision rules that produced non-select classifications. Each tree node also had an associated misclassification rate. This rate was used to calculate the tree node's confidence level for its prediction. Figure 2 shows a generated rule of one of the select nodes from Model 22. The graph on the right is a simple two-dimensional view of how C&RT works. Each of the partnership returns (represented by the red dot) was plotted onto this heterogeneous space using two fields, total deduction in the x-axis and ordinary income in the y-axis. In this example, the first split (Orthogonal line) using the Gini algorithm was placed between total deductions greater than and equal to $566,193 and total deduction less than $566,193. Then another split was created between ordinary income greater than and equal to -$422,056 and ordinary income less than -$422,056. These splits have created homogenous populations or segments that each of the cases to fall into (The graph does not show the other two splits created on gross profit and bad debt). The sub-group in the upper right corner with the cases labeled 'S' for selects and 'N' for non-selects graphically represents how the cases segmented into this homogeneous group thereby creating a decision rule or tree node. Darwin automatically calculates the misclassification cost related to the cases with the lowest representation of values for the segment, which in this case are the non-select (N) cases. The misclassification cost was calculated by dividing the number of misclassified record (2 'N' cases) by the total record (10 cases, 8 'S' and 2 'N' cases). Therefore, this tree node from Model 22 would be accurate in identifying select cases 80% of the time.

Figure 3 How The Model was created

Darwin: How the model was created?

```
[ Model  22)
Total records: 10 (0.0176991)
Target records: 8 (0.0253968) ]

    IF   TOT_DEDU > 566193  AND
    ORD_INC > -422056  AND
    GR_PRFT_ <= 1.69724e+07  AND
BAD_DEBT > 32651.5
THEN  RANK3 = Select
WITH  misclassification cost = 0.2
```



VI. Model Application

The completed model was applied to the LMSB BRTF for processing years 2000 and 2001. Each record of the BRTF was identified as a "select" (should be examined) or "non-select" (accepted as filed). A confidence level was determined for each record. The confidence level indicates the likelihood for the Model 22 to accurately predict if a return will be a "select". For example, partnership returns scored as select at the 90% confidence level indicates that 90 out of every 100 returns would be correctly classified and 10 would be misclassified. In PY 2000, there were 55,288 partnership returns filed. The 1,760 sample returns were excluded from the file before running the model. In PY 2001, there were 60,142 returns. Results are shown in Table 3 for all confidence levels (1-100%).

Table 3 All Partnership Returns

	PY 2000	PY 2001	% Change
Select	14,449	18,025	25%
Non-Select	39,079	42,117	8%
Total	53,528	60,142	12%

A. Comparison of PY 2000 and PY 2001 Selects at 70% to 100% Confidence Levels

Returns identified as "selects" with a confidence level of 70% or higher were analyzed. The number and percentage of selects for PY 2000 were compared to the number and percentage of selects for PY 2001. [The differences between the two processing years with or without the 458 sample selects were not significant, therefore, no adjustment were made to account for these returns.]

Tables 4 through 6 show the results from applying the Model 22 to PY 2000 and PY 2001 returns. The numbers reflected on the Tables indicate the number of returns with good audit potential, selects (as defined through manual classification). The number of selects has increased both in absolute numbers and as a percentage of the partnership population from the PY 2000 number. The overall number of returns selected was 23.4% above the PY 2000 select. The number of returns with a confidence level of 80-89% increased significantly compared to the number of returns with a confidence level of 70-79%.

Table 4 PY 2000 Number and Percentage of Selects by Confidence Levels

Asset Class	Number of Returns Filed*	Number of selects by Confidence Levels**				Percentage of Selects to Number of Returns
		70-79%	80-89%	90%+	Total	
$10-14.9M	16581	1824	244	91	2159	13.0%
$15-49.9M	24579	4058	431	100	4589	18.7%
$50-99.9M	5759	1444	72	313	1829	31.8%
$100-249M	3909	1147	36	426	1609	41.2%
$>=250	2700	857	34	35	926	34.3%
Total	53528	9330	817	965	11112	20.8%

*****The Partnership population in PY00 was 55,288 (TC 150 only).
**The number of selects represent the model selects and it did not include the classifier selects (458 returns)

Table 5 PY 2001 Number and Percentage of Selects by Confidence Levels

Asset Classes	Number of Returns Filed	Number of selects by Confidence Levels				Percentage of Selects to Number Returns
		70-79%	80-89%	90%+	Total	
$10-14.9M	18553	2112	591	101	2804	15.1%
$15-49.9M	27335	4441	869	116	5426	19.8%
$50-99.9M	6529	1316	579	398	2293	35.1%
$100-249M	4534	1270	181	526	1977	43.6%
$>=250	3191	1093	76	46	1215	38.1%
Total	60142	10232	2296	1187	13715	22.8%

VII. Model Tests

Several tests were conducted to determine how well the model predicts. Although these tests may not accurately reflect the validity of the model, they serve as an indication of how well the model predicts productive returns (selects). A more systematic test to validate the efficacy of the Model 22 will be conducted when the actual audit results from the model selected returns become available.

A. Partnership Technical Adviser vs. Classifier

In the review of the classifiers' determinations, the Partnership Technical Advisers reclassified 177 "selects" as "accepts." At 80% confidence level the Model agreed with the Technical Advisors on 166 of the 177 returns (94%). See table 7. This is an *indication* that the model is accurately making the select/accept determination.

Table 6 Analyses of the Partnership Technical Advisers, Classifier Scoring vs. Model

		Model	
		Select	**Non-Select**
Partnership TA	Non-Select	11	166

B. Comparison of Operating Partnership to Pass Through Partnerships

A separate analysis on operating partnerships (trade or business) and pass through partnerships (Schedule K only) was conducted to determine if there were significant differences between their number of selects and non-selects. A pass through partnership for the purpose of this analysis is defined as returns with no entries on the face of the Form 1065. Tables 8-10 below show the results of analysis of "selects" at the 70% or higher confidence level. The higher select rate for operating partnerships is consistent with knowledge or compliance risk and audit potential.

Table 7 PY 2000 Trade or Business

Type of Business	Number of Returns Filed	Selects	Non-Selects	Percentage of Selects to Number of Returns at 70%+ Confidence Level
Trade or Business	24,479	10,434	14,045	42.6%
Schedule K	29,409	678	28,371	2.28%
Total	53,528	11,112	42,416	20.8%

Table 8 PY 2001 Trade or Business

	Number of Returns Filed	Selects	Non-Selects	Percentage of Selects to Number of Returns at 70%+ Confidence Level
Trade or Business	27,927	12,485	15,442	44.71%
Schedule K	32,215	1,230	30,985	3.82%
Total	60,142	13,715	46,427	22.80%

Table 9 Number and Percentage Difference between PY 2000 and PY 2001

	Number of Returns Filed	Selects	Non-Selects	Percentage of change of selects
Trade or Business	3448	2051	1397	19.6%
Schedule K	2,582	552	2614	81.4%
Total	6030	2603	4011	23.4%

C. Comparison of Model 22 and DIF

DIF is currently the workload selection system that identifies potentially non-compliant partnership returns. The Tables that follow show the comparison of the number of returns predicted to be selects and non-selects using Model 22 (at the 70% confidence level) and DIF (score of 800 or higher). Shaded cells represent DIF and model agreement.

Table 10 Model 22 and DIF Comparison Using PY 2000 BRTF

Model 22

DIF	Select	Non-Select	Total	% DIF
Select	6,061	3,960	10,021	*19%*
Non-Select	5,051	38,856	43,507	*81%*
Total	11,112	42,416	53,528	*100%*
% Model 22	*21%*	*79%*	*100%*	

- DIF agreed with 6,061 of 11,112 Model 22 selects (54%)
- Model 22 agreed with 6,061 of 10,021 DIF selects (60%)
- Model 22 identified 11% more selects than DIF

Table 11 DIF and Model 22 Comparison Using PY 2001 BRTF

		Model 22			
		Select	**Non-Select**	**Total**	**% DIF**
DIF	Select	7,074	4,817	11,891	19.8%
	Non-Select	6,641	41,610	48,251	80.2%
	Total	13,715	46,427	60,142	100%
	% Model 22	22.8%	77.2%	100%	

- DIF agreed on 7,074 of 13,715 Model 22 selects (52%)
- Model 22 agreed on 7,074 of 11,891 DIF selects (60%)
- Model 22 identified 15% more selects than DIF

The lack of convergence between Model 22 and DIF is not surprising since, as explained above, there are reasons to suspect the accuracy of DIF for LMSB partnership returns. PY 2001 returns were ordered based on DIF score and other criteria. From this population, "selects" from Model 22 at the 70% confidence level were then identified. When manual classification was started, it was found that the select rate was so high that further classification was unnecessary. Returns identified by Model 22 were sent directly to the field for examination. Again, this experience is evidence that Model 22 does accurately identify returns that would be selected for examination if classified manually.

A test is now being conducted to score prior year returns and compare Model 22 results with actual audit results. As mentioned above, there is some concern that past audit results are not a good indicator of actual relative non-compliance this comparison should provide more information about the model's effectiveness. A final test will be conducted using results of examinations of returns selected by Model 22. Because of the time necessary to complete the examinations, this cannot be started for 12-18 months.

VIII. Findings and Limitations

A. Findings

Data mining can be used to develop models capable of predicting potentially non-compliant partnership returns. Oracle's Darwin software is a very useful tool for this kind of application. The model application and model tests supports the applicability to the LMSB partnership population.

A comparison between PY2000 and PY 2001 filings reflect an increase in the number of filings, but the number of "selects" has increased both in absolute numbers and as a percentage of the partnership population. This may reflect an increase in potential compliance risk in the LMSB partnership return population.

Model 22 and the current DIF system agree on "selects" only a little more than 50% of the time. Using Model 22 in combination with DIF may provide the best returns for examination.

B. Limitations

Model 22 is limited to the LMSB partnership population. The ranking used in the model development is dependent upon the accuracy of the classifiers' judgment. Model 22 was not developed to predict audit results since audit results from the sample would not be available for several years. Model 22 was built on the transcribed data from PY 2000 BRTF and any additional data transcription may impact the results of the model.

IX. Conclusion

LMSB West in concert with Oracle Corporation successfully achieved the research objective of developing and delivering an automated scoring model that identifies potentially non-compliant returns replicating judgments made by experience classifiers. Complete validation of Model 22 will not be possible until audit results from returns selected using the model are received and analyzed.

X. Recommendations

Implementation of the following recommendations should result in reduced IRS classification costs, improved examination results and reduced burden to compliant taxpayers.

- Use Model 22 as a further validation of DIF in order to increase the probability of selecting potentially non-compliant returns.
- Recalibrate Model 22 annually. This will incorporate any tax law changes, changes in taxpayer behavior and transcription changes. It is critical to update the model annually until the true refreshment period is determined.
- Evaluate Model 22 using historical audit results and results of returns selected using the model.
- Develop risk assessment and scoring models across the entire partnership population.
- Create a new model using Neural Network techniques that will predict the dollar value of each partnership select

XI. Cost and Benefits

The total cost for development and initial classification for this project is $306,000. This includes travel, staffing and contractor costs as shown in Tables 14 to 16.The staffing cost includes the actual salary of the employees working on this project plus 25% benefit rate (Table 14).

Table 12 Staffing Costs

Item	Staff Days	Total Hours	Cost per Hour	Total
LMSB Tax Audit Classification Tax Audit Specialists LMSB Research West	670	5,360	$36.56	$196,000

Table 13 Travel Costs

Item	Travel Days	Cost Per Day	Total
Travel To and From Classification Sites	300	$200	$60,000

Table 14 Outside Contractor Costs

Item	Work Day	Cost Per Day	Total
Work performed by outside contractor – Oracle Corporation The total (staffing and travel)	18.3	$2,738	$50,000

Statement of benefits: The model developed by LMSB Research and Oracle Corporation has the ability to classify the entire LMSB partnership return population. Use of the automated process:

- Saves classification time and travel costs by eliminating the need for a centralized manual classification and permits use of these resources on casework
- Reduces overall cycle time (filing to resolution) by delivery work to the field sooner
- Improves examination results and reduces the burden to taxpayers of unnecessary examinations

Actual savings/benefits will not be determined until such time as audit results from Model 22 selected returns have been analyzed.

XII. Milestones

The table below reflects the timeline for completing this project.

Task	Start Date	Finish Date
Hold Planning Meeting	4/18/2001	4/20/2001
Develop Sampling Plan	5/1/2001	5/6/2001
Develop A Scoring Model	12/10/2001	1/30/2001
Select Sample Returns for Classification	5/1/2001	5/6/2001
Develop Classification Check sheet	5/1/2001	5/16/2001
Develop Classifier training and Quality Assurance Plan	6/1/2001	6/6/2001
Manually Classify Returns	6/24/2001	9/22/2001
Meet with Consultant to discuss development of scoring model	7/16/2001	7/18/2001
Develop and Perfect Dataset	6/24/2001	12/7/2001
Deliver dataset to Oracle	12/10/2001	12/11/2001
Contractor to Develop Scoring Model based issues identified	12/10/2001	1/30/2002
Assign returns for audit that meet exam and workload criteria	8/15/2001	10/14/2001
Test the Scoring Model	2/1/2002	In Process
Develop Data File containing AIMS and BRTF data	2/1/2002	4/29/2002
Validate model against AIMS	4/29/2002	5/15/2002

XIII. Privacy/Security

The participants of the Partnership Project limited data to that necessary for the completing the project. Public and official access to the information was controlled. The data given to Oracle was stripped of any information that could identify taxpayers.

Security requirements were based on the Computer Security Act of 1987 and Office of Management and Budget Circular A-130, Appendices A7B. The security of the data used in this study and the privacy of the taxpayers was carefully safeguarded at all times. Physical security measures included a locked, secured office. Magnetic media was stored in locked cabinets. Printouts of sensitive data was stored in locked cabinets or shredded.

Data security was accomplished via Windows NT operating system. Systems were password protected, users were profiled for authorized use, and individual audit trails were generated and reviewed.

LMSB Research applied appropriate information and record keeping practices to ensure protection of all taxpayers. Oracle Corporation and its representative abided by the disclosure requirement outlined in the 'Statement of Work'.